Speedup of GEMM on A100 and RTX 4090 (FP16) cuBLAS BitBLAS-RTX4090 CUTLASS-A100 BitBLAS-A100 1.4 CUTLASS-RTX4090 AS 7**B**n₂ 8.0 Speedup o.4 0.2 0.0 М2 Shapes from LLM